arXiv:1301.3683v1 [math.OC] 16 Jan 2013

Convex Variational Image Restoration with Histogram Priors

Paul Swoboda and Christoph Schnörr

January 17, 2013

Abstract

We present a novel variational approach to image restoration (e.g., denoising, inpainting, labeling) that enables to complement established variational approaches with a histogram-based prior enforcing closeness of the solution to some given empirical measure. By minimizing a single objective function, the approach utilizes simultaneously two quite different sources of information for restoration: spatial context in terms of some smoothness prior and non-spatial statistics in terms of the novel prior utilizing the Wasserstein distance between probability measures. We study the combination of the functional lifting technique with two different relaxations of the histogram prior and derive a jointly convex variational approach. Mathematical equivalence of both relaxations and optimality certificates are established. Additionally, we present an efficient algorithmic scheme for the numerical treatment of the presented model. Experiments using the basic total-variation based denoising approach as a case study demonstrate our novel regularization approach.

1 Introduction

A broad range of powerful variational approaches to low-level image analysis tasks exist, like image denoising, image inpainting or image labeling [9], [8]. It is not straightforward however to incorporate *directly* into the restoration process statistical prior knowledge about the image class at hand. Particularly, handling global statistics as part of a single variational approach has not been considered so far.

In the present paper, we introduce a class of variational approaches of the form

$$\inf_u R(u) + F(u) + W(\mu^u, \mu^0), \tag{1}$$

where $R(u)+F(u)$ is any basic variational approach consisting of a regularization term $R(u)$, a data fidelity term $F(u)$ and $W(\mu^u, \mu^0)$ denotes the histogram prior in terms of the Wasserstein distance between the histogram corresponding to the minimizing function u to be determined and some given histogram μ^0. We require $R(u)$ to be convex which holds naturally in the case of denoising. As

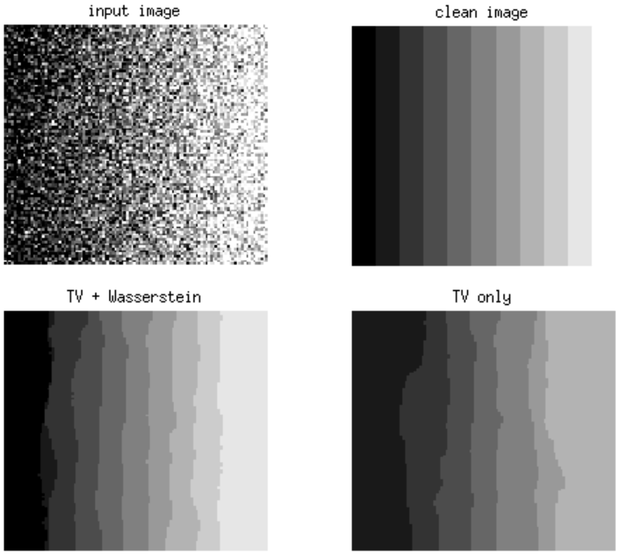


Figure 1: Denoising experiment of a noisy image (upper row, left side) taking into account statistical prior information through convex optimization (lower row, left side) infers the correct image structure and outperforms hand-tuned established variational restoration (lower row, right side). Enforcing global image statistics to be similar to those of the clean image (upper row, right side) gives our approach an advantage over methods not taking such information into account.

a case study, we adopt for $R(u) = TV(u)$, the Total Variation, see [2], and $F(u) = \int_\Omega f(u(x), x)dx$, where f can also be a nonconvex function. The basic ROF denoising approach of [16] is included in this approach with $f(u(x), x) = \|u(x) - u_0(x)\|^2$, where u_0 is the image to be denoised.

Note that minimizing the first term in (1) entails spatial regularization whereas the third Wasserstein term utilizes statistical information that is not spatially indexed in any way. Combining such different sources of information into a single variational approach has not been studied so far, to our knowledge. As an illustration, consider the academical example in figure 1. Knowing the greyvalue distribution of the original image helps us in regularizing the noisy input image. We tackle the corresponding main difficulty in two different, mathematically plausible ways: by convex relaxations of (1) in order to obtain a computationally tractable approach. Comparing these two relaxations – one may be tighter than the other one – reveals somewhat surprisingly mathematical equivalence. Preliminary numerical experiments demonstrate that the relaxation seems to be tight enough so as to bias effectively variational restoration towards given statistical prior information.

Our paper is organized as follows. After briefly reviewing relevant work in section 2, we state our approach and the related variational problem in section 3. Sections 4 and 5 provide two apparently quite different relaxation approaches that enable to solve approximately the variational problem by convex optimization. Equivalence of these two relaxations is established in section 6. We conclude with an efficient algorithm for computing a solution to the relaxation and show numerical examples in section 8 that demonstrate basic properties of our approach.

2 Prior Work

Image regularization by variational methods is a powerful and commonly used tool for denoising, inpainting, labeling and many other applications. As a case study in connection with (1), we consider one of the most used approaches for denoising, namely the Rudin, Osher and Fatemi (ROF) model from [16]:

$$\min_{u \in \mathrm{BV}} \|u - u_0\|^2 + \alpha \mathrm{TV}(u), \tag{2}$$

where u_0 is the input image, TV denotes the Total Variation and BV is the space of functions of bounded variation. The minimization problem (2) is convex and can be solved to a global optimum efficiently by various sparse first-order algorithms even for large problem sizes, e.g. by Primal-Dual methods, see [4], or other algorithms for nonsmooth convex optimization like [13], which we will use in our paper.

We can also use more general data terms instead of the quadratic term in (2). For example in [11] it is shown how the data term can be replaced by an (almost) arbitrary nonconvex function $\int_\Omega f(u(x), x)dx$. Still this data function is local and does not take into account global statistics.

In the case that some prior knowledge is encoded as a histogram, the Wasserstein distance and the associated Optimal Transport is a suitable choice for penalizing deviance from prior knowledge in a reasonable way. More generally the Wasserstein distance can be used as a distance on histograms over arbitrary metricized spaces.

Regarding the Wasserstein distance and the theory of Optimal Transport, we refer to the in-depth treatise [17]. Optimal Transport is well-known as Earth Mover's distance in image processing and computer vision [15] and is used for content-based image retrieval. Further recent applications include [5] and [10] in connection with segmentation and [6] for texture models.

Our histogram-based prior detailed below, employing the Wasserstein distance, and its use as part of a single and general variational approach through convex relaxation, appears to be novel.

3 Problem and Mathematical Background

We introduce the original non-convex model, consider different ways to write the Wasserstein distance and introduce the functional lifting technique for rewriting the resulting optimization problem to make it amenable for optimization.

3.1 Problem Statement

For an image domain $\Omega \subset \mathbb{R}^2$, e.g. $\Omega = [0,1]^2$ and $u : \Omega \to [0,1]$, consider the following integral over Dirac measures:

$$\mu^u = \frac{1}{|\Omega|} \int_\Omega \delta_{u(x)} dx, \tag{3}$$

where

$$\delta_y(A) = \begin{cases} 1 & , y \in A \\ 0 & , y \notin A \end{cases} \tag{4}$$

is the Dirac measure. (3) is the grey-value histogram of the image u, i.e.

$$\mu^u(A) = \frac{1}{\Omega} |u^{-1}(A)|, \tag{5}$$

where $A \subset [0,1]$ and $|B|$ is the Lebesgue measure of a set $B \in \Omega$.

We would like to minimize the energy function

$$\min_{u:\Omega \to [0,1]} E(u) = \int_\Omega f(u(x), x) dx + TV(u) + W(\mu^u, \mu^0). \tag{6}$$

$TV(u)$ is the Total Variation

$$TV(u) = \sup \left\{ \int_\Omega u(x) \cdot \operatorname{div} \phi(x) dx \ : \ \phi \in C_c^1(\Omega, \mathbb{R}^2), \|\phi\|_{L^\infty} \leq 1 \right\} \tag{7}$$

see [2] for more details. $f : [0,1] \times \Omega \to \mathbb{R}$ is a measurable but apart from that arbitrary fidelity function, and W is the Wasserstein distance

$$W(\mu, \nu) = \inf_{\pi \in \mathcal{M}(\mu,\nu)} \int_{[0,1]\times[0,1]} c \ d\pi. \tag{8}$$

$c : [0,1] \times [0,1] \to \mathbb{R}$ is the cost function for the Wasserstein distance, for example $c(x,y) = |x-y|^p$ with $p \geq 1$. The space of transport plans is

$$\mathcal{M}(\mu, \nu) = \{\pi \in \mathcal{P}([0,1]\times[0,1]) : \begin{array}{l} \pi(A \times [0,1]) = \mu(A) \\ \pi([0,1] \times B) = \nu(B) \end{array} \quad \forall A, B \text{ measurable}\}, \tag{9}$$

where $\mathcal{P}([0,1] \times [0,1])$ is the space of all probability measures defined on the Borel-σ-Algebra over $[0,1]\times[0,1]$. By Theorem 4.1 in [17] there exists a measure which minimizes (8) and it is called the optimal transport plan. (8) is a linear minimization problem subject to linear constraints and therefore convex. Note however that energy (6) is not convex.

By minimizing (6) we obtain a solution u which remains faithful to the data by the fidelity term f, is spatial coherent by the Total Variation term and has global grey-value statistics similar to μ_0.

3.2 The Wasserstein Distance and its Dual

We reformulate energy (6) to make it more amenable for numerical computation by introducing another way to obtain the Wasserstein distance. For this we recall Theorem 5.10 in [17], which states that the following dual Kantorovich formulation equals the Wasserstein distance:

$$W(\mu^u, \mu^0) = \sup_{\{(\psi,\psi') \,:\, \psi(x)-\psi'(y) \leq c(x,y)\}} \int_0^1 \psi d\mu^u - \int_0^1 \psi' d\mu^0. \tag{10}$$

Define therefore

$$E(u, \psi, \psi') = \int_\Omega f(u(x), x)dx + TV(u) + \int_0^1 \psi d\mu^u - \int_0^1 \psi' d\mu^0 \tag{11}$$

and let

$$C = \mathrm{BV}(\Omega, [0,1]) \tag{12}$$

be the space of functions of bounded variation with domain Ω and range $[0,1]$ and

$$D = \left\{ \psi, \psi' : [0,1] \to \mathbb{R} \text{ s.t. } \begin{array}{c} \psi(x) - \psi'(y) \leq c(x,y) \ \forall x, y \in [0,1] \\ \psi, \psi' \text{ measurable} \end{array} \right\}. \tag{13}$$

It follows from (10) with the above definitions that

$$\min_{u \in C} E(u) = \min_{u \in C} \max_{(\psi,\psi') \in D} E(u, \psi, \psi') \tag{14}$$

3.3 Functional Lifting

While the Wasserstein distance (8) is convex in both of its arguments, see Theorem 4.8 in [17], (6) is not convex due to the nonconvex transformation $u \mapsto \mu^u$ in the first argument of the Wasserstein term and the possible nonconvexity of f. To overcome the nonconvexity of both the data term and the transformation in the first argument of the Wasserstein distance we lift the function u. Instead of u we consider a function ϕ below whose domain is one dimension larger. This extra dimension represents the range of u and allows us both to linearize the fidelity term and to convexify the Wasserstein distance. This technique, known as functional lifting or the calibration method, was introduced in [1] and is commonly used in many optimization problems.

Let

$$C' = \{\phi : \Omega \times [0,1] \to \{0,1\} : \ \phi(\cdot, 0) = 1, \ \phi(\cdot, 1) = 0, \ \partial_2 \phi \leq 0, \ \phi \in BV\}. \tag{15}$$

Every function $u \in C$ corresponds uniquely to a function $\phi \in C'$ via the relation

$$u(i,j) = y \quad \Leftrightarrow \quad -\partial_2 \phi(x, \cdot) = \delta_y. \tag{16}$$

Also for such a pair (u, ϕ) we have the relation

$$\mu^u = \mu^\phi = \frac{1}{|\Omega|}\int_\Omega |\partial_2\phi(x,\cdot)|dx = \frac{1}{|\Omega|}\int_\Omega -\partial_2\phi(x,\cdot)dx. \tag{17}$$

In contrast to the transformation $u \mapsto \mu^u$, $\phi \mapsto \mu^\phi$ is linear.

Consider the energy

$$\begin{aligned} E'(\phi,\psi,\psi') &= \textstyle\int_0^1 TV(\phi(\cdot,\gamma))d\gamma + \int_0^1\int_\Omega |\partial_2\phi(x,\gamma)|f(\gamma,x)dxd\gamma \\ &\quad + \textstyle\int_0^1 \psi d\mu^\phi - \int_0^1 \psi' d\mu^0. \end{aligned} \tag{18}$$

For a pair (u, ϕ) as above the identity

$$E(u,\psi,\psi') = E'(\phi,\psi,\psi') \tag{19}$$

holds true by the coarea formula 3.40, see [2]. Consequently, we have

$$\min_{u\in C}\max_{(\psi,\psi')\in D} E(u,\psi,\psi') = \min_{\phi\in C'}\max_{(\psi,\psi')\in D} E'(\phi,\psi,\psi'). \tag{20}$$

Note that E' is convex in ϕ and concave in (ψ,ψ').

Remark 1. *As discussed in Section 1, we merely consider total variation based regularization as a case study, but this restriction is not necessary. More general regularizers can be used as well. In the present paper however, we rather focus on the novel prior based on the Wasserstein distance.*

4 Relaxation as a Convex/Concave Saddle Point Problem

Optimizing energy (6) is not tractable, as it is a nonconvex problem. Also solving (20) is not tractable, as the set C' is nonconvex. The latter can be overcome by considering the convex hull of C', which leads to a relaxation as a convex/concave saddle point problem for the minimization problem (6), which is solvable computationally.

Proposition 1. *Let*

$$C'' = \{\phi : \Omega\times[0,1] \to [0,1] \, : \; \phi(\cdot,0)=1, \; \phi(\cdot,1)=0, \; \partial_2\phi \le 0, \; \phi \in BV\} \tag{21}$$

Then E' is convex/concave and

$$\min_{u\in C} E(u) \ge \min_{\phi\in C''}\sup_{(\psi,\psi')\in D} E'(\phi,\psi,\psi'). \tag{22}$$

If

$$\min_{u\in C}\max_{(\psi,\psi')\in D} E(u,\psi,\psi') = \max_{(\psi,\psi')\in D}\min_{u\in C} E(u,\psi,\psi') \tag{23}$$

holds, then the above relaxation is exact.

Proof. Note that C'' is a convex set, in particular it is the convex hull of C'. E' is also convex in ϕ, therefore the right side of (22) is a convex/concave saddle point problem. For fixed (ψ, ψ') we have the following equality:

$$\min_{u \in C} E(u, \psi, \psi') = \min_{\phi \in C''} E'(\phi, \psi, \psi'), \tag{24}$$

which is proved in [11]. Thus

$$\begin{aligned} \min\nolimits_{u\in C} E(u) &= \min\nolimits_{u\in C} \max\nolimits_{(\psi,\psi')\in D} E(u,\psi,\psi') \\ &\overset{(*)}{\geq} \max\nolimits_{(\psi,\psi')\in D} \min\nolimits_{u\in C} E(u,\psi,\psi') \\ &\overset{(**)}{=} \max\nolimits_{(\psi,\psi')\in D} \min\nolimits_{\phi\in C''} E'(\phi,\psi,\psi'), \end{aligned} \tag{25}$$

where $(*)$ is always fulfilled for minimax problems and $(**)$ is a consequence of (24). This proves (22). If (23) holds, then $(*)$ above is actually an equality and the relaxation is exact. □

5 Relaxation with Hoeffding-Fréchet Bounds

A second relaxation can be constructed by using the primal formulation (8) of the Wasserstein distance and enforcing the marginals of the distribution function of the transport plan to be μ^ϕ and μ^0 by the Hoeffding-Fréchet Bounds:

Theorem 1. *Let F_1, F_2 be two real distribution functions (d.f.s) and F a d.f. on $\mathbb{R}^2$. Then F has marginals F_1, F_2, if and only if*

$$(F_1(x_1) + F_2(x_2) - 1)_+ \leq F(x_1, x_2) \leq \min\{F_1(x_1), F_2(x_2)\} \tag{26}$$

Proof. see Theorem 3.1.1 in [12] □

By (8) the Wasserstein Distance with marginal d.f.s F_1, F_2 can be computed by solving the optimal transport problem and we arrive at the formulation

$$W(dF_1, dF_2) = \min_F \int_{\mathbb{R}^2} c \; dF, \quad \text{s.t. } F \text{ respects the Hoeffding-Fréchet Bounds,} \tag{27}$$

where dF_i shall denote the measure associated to the d.f. F_i, $i = 1, 2$.

Using again the functional lifting technique of [11], the Hoeffding-Fréchet-Bounds and the representation of the Wasserstein distance (27), we arrive at the following relaxation.

$$\begin{aligned} \min\nolimits_{\phi,F} \quad & \textstyle\int_0^1 TV(\phi(\cdot,\gamma)d\gamma + \int_0^1 \int_\Omega |\partial_2 \phi(x,\gamma)| \cdot f(\gamma,x) dx d\gamma + \int_{\mathbb{R}^2} c \; dF, \\ s.t. \quad & F_\phi(x) = \tfrac{1}{|\Omega|} \textstyle\int_\Omega \int_0^x |\partial_2 \phi(x,\gamma)| d\gamma dx, \\ & F_{\mu^0}(x) = \textstyle\int_0^x d\mu^0, \\ & F_\phi(x_1) + F_{\mu^0}(x_2) - 1 \leq F(x_1,x_2) \leq \min\{F_\phi(x_1), F_{\mu^0}(x_2)\} \\ & \phi \in C'' \end{aligned} \tag{28}$$

(28) is a relaxation of (6). Just set

$$\phi(i,j,\gamma) = \begin{cases} 1, & u(i,j) < \gamma \\ 0, & u(i,j) \geq \gamma \end{cases}$$

and let F be the d.f. of the optimal transport measure with marginals μ^u and μ^0.

Remark 2. *It is interesting to know, when relaxation* (28) *is exact. By the coarea formula [18] we know that*

$$\begin{array}{rl} & \int_0^1 TV(\phi(\cdot,\gamma))d\gamma + \int_0^1 \int_\Omega |\partial_2 \phi(x,\gamma)| \cdot f(\gamma,x) dx d\gamma \\ = & \int_0^1 TV(u_\alpha) d\alpha + \int_0^1 \int_\Omega f(u_\alpha(x),x) dx d\alpha, \end{array} \tag{29}$$

where u_α coresponds to $\phi_\alpha = \mathbb{1}_{\{\phi > \alpha\}}$ via relation (16)*. However such a formula does not generally hold for the optimal transport: Let*

$$\phi_\alpha = \mathbb{1}_{\{\phi > \alpha\}} \tag{30}$$

and let F_α be the d.f. of the optimal coupling with marginal d.f.s F_{ϕ_α} and F_{μ^0}. Then

$$F = \int_0^1 F_\alpha \; d\alpha \tag{31}$$

has marginal d.f.s $\int_0^1 F_{\phi_\alpha} d\alpha$ and F_{μ^0}, but it may not be optimal.

A simple example for inexactness can be constructed as follows: Let the data term be $d = 0$ and let $\mu_0 = \frac{1}{2}(\delta_0 + \delta_1)$ and let the cost for the Wasserstein distance be $c(x,y) = \lambda|x-y|$. Every constant function with $u(i,j) = \text{const} \in [0,1]$ will be a minimizer if λ is small enough. The objective value will be $\frac{\lambda}{2}$. But relaxation (28) *is inexact in this situation: Choose $\phi(x,\gamma) = \frac{1}{2} \quad \forall \gamma \in (0,1)$ and the relaxed objective value will be* 0*.*

6 Relationship between the two Relaxations

Both relaxations from Sections 4 and 5 seem to be plausible but seemingly different relaxations. Their different nature reveals itself also in the conditions for which exactness was established. While the condition in Proposition 1 is dependent the gap introduced by interchanging the minimum and maximum operation, relaxation (28) is exact if a coarea formula holds for the optimal solution. It turns out, that both equations are equivalent however, hence both optimality conditions derived in sections 4 and 5 can be used to ensure exactness of a solution to either one of the relaxed minimization problems.

Theorem 2. *The optima of the two relaxations* (22) *and* (28) *are equal.*

Proof. It is a well known fact that

$$\min_{x \in X} \max_{y \in Y} \langle Kx, y \rangle + G(x) - F^*(y) \tag{32}$$

and

$$\min_{x\in X} F(Kx)+G(x) \tag{33}$$

are equivalent, where $G : X \to [0,\infty]$ and $F^* : Y \to [0,\infty]$ are proper, convex, lower-semicontinuous (l.s.c.) functions, F^* is the convex conjugate of the convex l.s.c. function F and X and Y are two real vector spaces, see [14] for details.

To apply the above result choose

$$G(\phi)=\int_0^1 TV(\phi(\cdot,\gamma))d\gamma+\int_0^1\int_\Omega |\partial_2\phi(x,\gamma)|\cdot f(\gamma,x)dxd\gamma+\iota_{C''}(\phi), \tag{34}$$

$$F^*(\psi,\psi')=\int_0^1 \psi' d\mu^0+\iota_D(\psi,\psi') \tag{35}$$

and

$$\begin{gathered} K : BV(\Omega\times[0,1]) \to \mathcal{M}([0,1])^2, \\ K(\phi)=(\mu^\phi,0) \end{gathered} \tag{36}$$

where C'' is defined by (21) and D by (13), $\iota_{C''}(\cdot)$ and $\iota_D(\cdot)$ denote the indicator functions of the sets C'' and D respectively and $\mathcal{M}([0,1])$ denotes the space of measures on $[0,1]$. (32) corresponds with the above choices to the saddle point relaxation (22).

Recall that $F=(F^*)^*$, i.e. F is the Legendre-Fenchel bidual of itself, see [14]. Hence, for positive measures μ,ν, the following holds true:

$$\begin{aligned} F(\mu,\nu) = {} & \sup_{\psi,\psi'}\{\textstyle\int_0^1 \psi d\mu-\int_0^1 \psi' d\nu - F^*(\psi,\psi')\} \\ = {} & \sup_{(\psi,\psi')\in D}\{\textstyle\int_0^1 \psi d\mu-\int_0^1 \psi' d\nu-\int_0^1 \psi' d\mu^0\} \\ = {} & \sigma_D(\mu,\nu+\mu^0) \\ \overset{(*)}{=} {} & \begin{cases} W(\mu,\mu^0) & ,\nu=0 \\ \infty & ,\text{otherwise} \end{cases} \end{aligned} \tag{37}$$

where $\sigma_A(x)=\sup_{a\in A}\langle a,x\rangle$ is the support function of the set A. To prove $(*)$, we invoke Theorem 5.10 in [17], which states that

$$\sigma_D(\mu,\nu)=\sup_{(\psi,\psi')\in D}\int_0^1 \psi d\mu - \int_0^1 \psi'\nu = \min_{\pi\in\Pi(\mu,\nu)}\int_{[0,1]^2} c(x,y)d\pi(x,y)=W(\mu,\nu), \tag{38}$$

and we have infinity for measures which do not have the same mass.

Thus (33) can be written as

$$\min_\phi G(\phi)+\sigma_D(-\frac{1}{|\Omega|}\int_\Omega \partial_2\phi(x,\cdot)dx,\mu^0). \tag{39}$$

This problem is relaxation (28), which concludes the proof. □

7 Optimization

We present five experiments and the numerical method used to compute them.

7.1 Implementation

First, we discretize the infinite dimensional set C'' and denote it by C''_d. Hence we consider only finitely many grey values, which an image can take. Second we discretize the image domain Ω to be $\{1, \ldots, n_1\} \times \{1, \ldots, n_2\}$ and use as the gradient operator forward differences.

We use the Generalized Forward-Backward Splitting algorithm as in [13] for solving the relaxation (22).

Algorithm 1: The Generalized Forward-Backward Splitting Algorithm

```
Data: $(z_i)_{i\in\{1,\ldots,k\}} \in \mathcal{H}$,
$(\omega_i)_{i\in\{1,\ldots,k\}}$, $s.t.$ $\sum_{i=1}^{k} \omega_i = 1$
Result: $x$ optimal for the energy $\langle f, x\rangle + \sum_{i=1}^{n} G_i(x)$
Initialize:
$x \leftarrow \sum_{i=1}^{k} \omega_i z_i$
$t \leftarrow 0$
Main iteration:
repeat
    for $i \leftarrow 1$ to $k$ do
        $z_i \leftarrow z_i + \mathrm{prox}_{\frac{1}{\omega_i} G_i}(2x - z_i - f) - x$
    end
    $x \leftarrow \sum_{i=1}^{n} \omega_i z_i$
    $t \leftarrow t + 1$
until convergence;
```

Here $\mathcal{H} = C''_d \times (C''_d)^2$. The second component of $\mathcal{H}$ holds the gradient of ϕ. The convex functions G_i are:

- $G_1(\phi, g) = \delta_{\{\nabla\phi = g\}}$
- $G_2(g) = \|g\|_1$
- $G_3(\phi) = \delta_{C''}(\phi)$
- $G_4(\phi) = W(\mu_\phi, \mu^0)$

G_1 and G_2 are splitting the TV-term, as is customary, see Section 6.1.3. in [13].

The Generalized Forward-Backward Splitting algorithm requires to compute efficiently the proximity operators

$$\mathrm{prox}_{G_i}(x) = \mathrm{argmin}_{x'} \frac{1}{2}\|x - x'\|^2 + G_i(x'). \tag{40}$$

prox_{G_1} is the shrinkage operator and prox_{G_2} is the projection onto the set $\{\nabla\phi = g\}$. The latter can be efficiently computed with Fourier transforms, see again [13].

prox_{G_3} is the projection onto the set of non-increasing sequences C''. To compute this projection, we employ the algorithm proposed in [3], Appendix D. It is trivially parallelisable and converges in a finite number of iterations.

Last, the proximity operator for the Wasserstein distance can be computed efficiently in some special cases, as discussed next.

7.2 Wasserstein Proximation for $c(x, y) = |x - y|$ by shrinkage

In general, computing the proximity operator for the Wasserstein distance can be expensive and requires solving a quadratic program. However, for the real line and convex costs, we can compute the proximity operator more efficiently. One algorithm for the cost function $c(x, y) = |x - y|$ is presented below.

The proximation for the weighted Wasserstein distance is

$$\mathrm{argmin}_\phi \frac{1}{2}\|\phi^0 - \phi\|_2^2 + \lambda W(\mu^\phi, \mu^0). \tag{41}$$

For the special case we consider here, there is a simple expression for the Wasserstein distance:

Proposition 2. *For two measures μ, ν on the real line and $c(x, y) = |x - y|$, the Wasserstein distance is*

$$W(\mu, \nu) = \int_R |F_{\mu^1}(x) - F_{\mu^1}(x)| dx \tag{42}$$

Proof. Let the space of 1-Lipschitz functions on the real line be denoted by

$$L = \{\psi : \mathbb{R} \to \mathbb{R} \ : \ |\psi(x) - \psi(y)| \leq |x - y| \ \forall x, y \in \mathbb{R}\} \tag{43}$$

The Kantorovich dual of the Wasserstein distance (10) on the real line with cost function $c(x, y) = |x - y|$ for two histograms μ^1 and μ^2 with d.f. F_{μ^1} and F_{μ^2} reads, by Remark 5.4. in [17], as follows:

$$\begin{array}{rl} & \max_{\{\psi \in L\}} \int_R \psi \ d\mu^1 - \int_R \psi \ d\mu^2 \\ = & \max_{\{\psi \in L\}} \int_R \psi(x) \partial_x F_{\mu^1}(x) \ - \int_R \psi(x) \partial_x F_{\mu^2}(x) \\ = & \max_{\{\psi \in L\}} \int_R \psi(x) \partial_x (F_{\mu^1}(x) \ - F_{\mu^2}(x)) \\ = & \max_{\{\psi \in L\}} \int_R \partial_x \psi(x) (F_{\mu^1}(x) - F_{\mu^2}(x)) \ dx \\ = & \max_{\{|\rho| \leq 1\}} \int_R \rho(x) (F_{\mu^1}(x) - F_{\mu^2}(x)) dx \\ = & \int_R |F_{\mu^1}(x) - F_{\mu^2}(x)| dx \end{array} \tag{44}$$

When applying integration by parts above, no boundary terms occur because $F_{\mu^1}(x) - F_{\mu^2}(x)$ vanishes asymptotically. □

Due to $\partial_x \phi(i, j, x) \leq 0$ and $\phi(x, 0) = 1$, we can also write $F_{\mu^\phi}(\gamma)$ as

$$F_{\mu^\phi}(\gamma) = \int_0^\gamma \frac{1}{|\Omega|} \int_\Omega |\partial_2 \phi(x, \eta)| dx d\eta = \frac{1}{|\Omega|} \int_\Omega 1 - \phi(x, \gamma) dx \tag{45}$$

Next we show how to analytically solve the proximity operator for the Wasserstein distance in the present case.

Proposition 3. *Given ϕ^0, $\lambda > 0$, the optimal $\tilde{\phi}$ for the proximity operator*

$$\tilde{\phi} = \mathrm{argmin}_\phi \frac{1}{2}\|\phi - \phi^0\|_2^2 + \lambda W(F_{\mu^\phi}, \mu^0) \tag{46}$$

is determined by

$$\tilde{\phi}(x,\gamma) = \phi(x,\gamma) + c_\gamma, \tag{47}$$

where

$$c_\gamma = \mathit{shrink}\left(-\frac{1}{|\Omega|}\int_\Omega \phi_0(x,\gamma)dx - F_{\mu^0}(\gamma) + 1, \frac{|\Omega|}{\lambda}\right) + \frac{1}{|\Omega|}\int_\Omega \phi_0(x,\gamma)dx + F_{\mu^0}(\gamma) - 1 \tag{48}$$

and shrink denotes the shrinkage operator defined componentwise by

$$\mathit{shrink}(a,\lambda)^i = (|a^i| - \lambda)_+ \cdot \operatorname{sign}(a^i) \tag{49}$$

for $a \in \mathbb{R}^n$, $\lambda > 0$.

Proof. By proposition 2 and the characterisation of F_{μ^ϕ} in (45), proximation (41) reads

$$\operatorname{argmin}_\phi \frac{1}{2}\|\phi^0 - \phi\|_2^2 + \lambda \int_\mathbb{R} |1 - \left(\frac{1}{|\Omega|}\int_\Omega \phi(x,\gamma)dx\right) - F_{\mu^0}(\gamma)|d\gamma. \tag{50}$$

Note that (50) is an independent optimization problem for each γ. Thus, for each γ we have to solve the problem

$$\operatorname{argmin}_{\phi(\cdot,\gamma)} \frac{1}{2}\|\phi^0(\cdot,\gamma) - \phi(\cdot,\gamma)\|_2^2 + \lambda|1 - \left(\frac{1}{|\Omega|}\int_\Omega \phi(x,\gamma)dx\right) - F_{\mu^0}(\gamma)|. \tag{51}$$

It can be easily verified that the solution to problem (51) is $\phi^0(\cdot,\gamma) + c_\gamma$, where $c_\gamma \in \mathbb{R}$ and

$$c_\gamma \in \operatorname{argmin}_{c\in\mathbb{R}} \frac{1}{2}|\Omega|c^2 + \lambda|\frac{1}{|\Omega|}\int_\Omega \phi^0(x,\gamma)dx + c + F_{\mu^0}(\gamma) - 1| \tag{52}$$

and hence

$$c_\gamma = \operatorname{shrink}\left(-\frac{1}{|\Omega|}\int_\Omega \phi_0(x,\gamma) - F_{\mu^0}(\gamma) + 1, \frac{|\Omega|}{\lambda}\right) + \frac{1}{|\Omega|}\int_\Omega \phi_0(x,\gamma)dx + F_{\mu^0}(\gamma) - 1. \tag{53}$$

□

Concluding, the cost for the Wasserstein proximal step is linear in the size of the input data. For the discretized problem a similar computation holds.

8 Numerical Experiments

We want to show experimentally

1. that computational results conform to the mathematical model,
2. that the convex relaxation is reasonable.

Note that we do not claim to achieve the best denoising or inpainting results and we do not wish to compete with other state-of-the-art methods here. We point out again that the Wasserstein distance can be used together with other variational approaches to enhance their performance, e.g. with nonlocal total variation based denoising, see [7].

In the **first experiment** we compare total variation denoising and total variation denoising with the Wasserstein term for incorporating prior knowledge. The data term is $f(x,s) = \lambda \cdot \|u_0(x) - s\|^2$, where u_0 is the noisy image in figure 1. The cost for the Wasserstein distance is $c(x,y) = \nu|x-y|,\ \nu > 0$. To ensure a fair comparison, the parameter λ for total variation regularization *without* prior knowledge was *hand-tuned in all experiments* to obtain best results. The histogram was chosen to match the noiseless image. See figure 1 for the results.

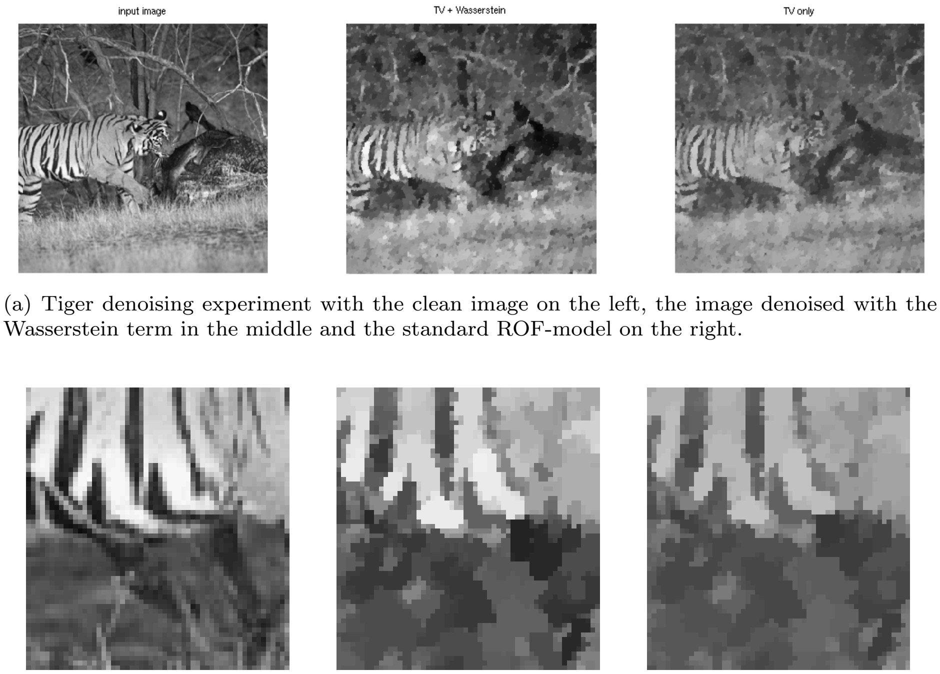
Note the trade-off one always has to make for pure total variation denoising: If one sets the regularization parameter λ high, the resulting grey-value histogram of the recovered image will be similar to the noisy input image and generally far away from the histogram of ground truth. By choosing lower data fidelity and higher regularization strength we may obtain a valid geometry of the image, however then the grey-value histogram tends to be peaked at one mode, as total variation penalizes scattered histograms and tries to draw the modes closer to each other, again letting the recovered grey-value histogram being different from the desired one.

The **second** experiment is a more serious denoising experiment. Notice that again pure total variation denoising does not preserve the white and black areas well, but makes them grey, while the approach with the Wasserstein distance preserves the contrast better, see figure 2.

In the **third experiment** we compare image inpainting with a total variation regularization term without prior knowledge and with prior knowledge, see figure 3 for the results. The region where the data term is zero is enclosed in the blue rectangle. Outside the blue rectangle we employ a quadratic data term as in the first exeriment. Total variation inpainting without prior knowledge does not produce the results we expected, as the total variation term is smallest, when the grey color fills most of the area enclosed by the blue rectangle. Heuristically, this is so because the total variation term weighs the boundary length multiplied by the difference between the grey value intensities, and a medium intensity minimizes this cost. Thus the TV-term tends to avoid interfaces, where high and low intensities meet, preferring smaller intensity changes, which can be achieved by interfaces with grey color on one side. Note that also the regularized image with the Wasserstein term lacks symmetry. This is also due to the behaviour of the TV-term described above.

In the **fourth** experiment we consider inpainting again. Yevgeni Khaldei, the photographer of the iconic picture shown on the left of figure 4 had to remove the second watch. Trying to inpaint the wrist with a TV-regularizer and a Wasserstein term results in the middle picture, while only using a TV-regularizer results in the right picture. Clearly using the Wasserstein term helps.

In the **fifth** experiment we have a different setup. The original image is on the left of figure 5. The histogram μ^0 was computed from a patch of clouds,



(a) Tiger denoising experiment with the clean image on the left, the image denoised with the Wasserstein term in the middle and the standard ROF-model on the right.

(b) Detailed view of the tiger denoising experiment revealing that contrast is better preserved when the Wasserstein term is used.

Figure 2: Tiger denoising experiment

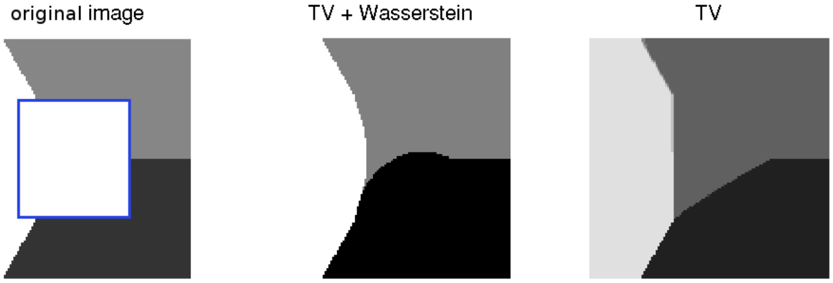


Figure 3: Inpainting experiment with the original image and the inpainting area enclosed in a blue rectangle on the left, the inpainting result with the Wasserstein term in the middle and the result where only the TV-regularizer is used on the right. By enforcing the three regions to have the same size with the Wasserstein term, we obtain a better result than with the Total Variation term alone.

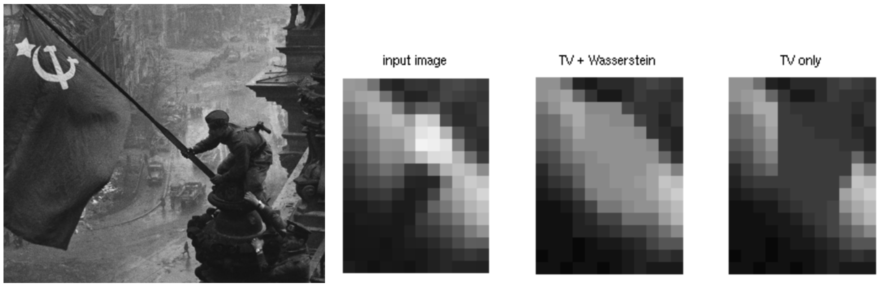


Figure 4: Here we want to inapint the area occupied by the watch of the soldier, see the second left image. Our approach, on the second right image gives better results again than the approach with TV alone.

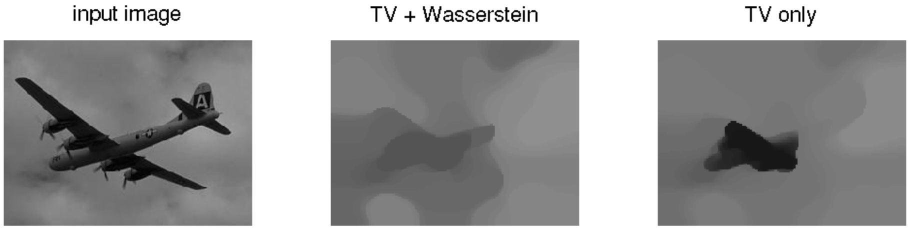


Figure 5: Unsupervised inpainting using empirical measures as priors. Objects not conforming to the prior statistics are removed *without* labeling image regions.

which did not include the plane. The data term is $f(x, y) = \lambda \min(|u_0(x) - y|^2, \alpha)$, where $\alpha > 0$ is a threshold, so the data term does not penalize great deviances from the input image too strongly. The Wasserstein term penalizes the image of the plane whose appearance differs from the prior statistics. The TV-regularizer is weighted weaker than in the previous examples, because we do not want to smooth the clouds.

Note that unlike in ordinary inpainting applications, we did not specify the location of the plane beforehand, but the algorithm figured it out on its own. The total variation term finally favors a smooth inpainting of the area occupied by the plane. In essence we have combined two different tasks: Finding out where the plane is and inpainting that area occupied by it. See figure 5 for results.

9 Conclusion and Outlook

We have presented in this paper a novel method for variational image regularization, which takes into account global statistical information in one model. By solving the relaxed nonconvex problem we obtain regularizd images which conform to some global image statistics, which sets our method apart from standard variational methods. Moreover the computational cost for the Wasserstein term we introduced is negligible, however our relaxation is not tight anymore as in models without the latter term. Still, the relaxation is reasonably tight.

Our future work will consider extensions of the present approach to multidimensional input data and related histograms, e.g. based on color, patches or gradient fields. The theory developed in this paper regarding the possible exactness of solutions does not carry over without modifications to such more complex settings. Moreover, it is equally important to find ways related to our present work to minimize such models efficiently.

References

[1] G. Alberti, G. Bouchitte, and G. Dal Maso. The calibration method for the Mumford-Shah functional and free-discontinuity problems. *Journal: Calc. Var. Partial Differential Equations*, 16:299–333, 2003.

[2] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of Bounded Variation and Free Discontinuity Problems (Oxford Mathematical Monographs).* Oxford University Press, USA, May 2000.

[3] A. Chambolle, D. Cremers, and T. Pock. A convex approach for computing minimal partitions. Technical report, Centre des Mathematiques Appliquees, Ecole Polytechnique, Palaiseau, Paris, France, 2008.

[4] A. Chambolle and T. Pock. A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.

[5] T. F. Chan, S. Esedoglu, and K. Ni. Histogram Based Segmentation Using Wasserstein Distances. In Fiorella Sgallari, Almerico Murli, and Nikos Paragios, editors, *SSVM*, volume 4485 of *Lecture Notes in Computer Science*, pages 697–708. Springer, 2007.

[6] S. Ferradans, G-S. Xia, G. Peyré, and J-F. Aujol. Optimal Transport Mixing of Gaussian Texture Models. Technical report, Preprint Hal-00662720, 2012.

[7] G. Gilboa and S. Osher. Nonlocal Operators with Applications to Image Processing. *Multiscale Modeling & Simulation*, 7(3):1005–1028, 2008.

[8] J. Lellmann and C. Schnörr. Continuous Multiclass Labeling Approaches and Algorithms. *SIAM J. Imag. Sci.*, 4(4):1049–1096, 2011.

[9] N. Paragios, Y. Chen, and O. Faugeras, editors. *The Handbook of Mathematical Models in Computer Vision.* Springer, 2006.

[10] G. Peyré, J. Fadili, and J. Rabin. Wasserstein Active Contours. Technical report, Preprint Hal-00593424, 2011.

[11] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. Global Solutions of Variational Models with Convex Regularization. *SIAM J. Imaging Sciences*, 3(4):1122–1145, 2010.

[12] S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems. Vol. I, Theory.* Springer-Verlag, New York, 1998.

[13] H. Raguet, J. Fadili, and G. Peyré. Generalized Forward-Backward Splitting. Technical report, Preprint Hal-00613637, 2011.

[14] R. T. Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics,. Princeton University Press, Princeton, princeton paperbacks edition, 1997.

[15] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40:2000, 2000.

[16] L. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268, 1992.

[17] C. Villani. *Optimal Transport: Old and New.* Grundlehren der mathematischen Wissenschaften. Springer, 1 edition, November 2008.

[18] W.P. Ziemer. *Weakly Differentiable Functions.* Springer, 1989.